THE RIMINI CENTRE FOR ECONOMIC ANALYSIS

WP 10-20

Theodore Panagiotidis
Department Of Economics, University Of Macedonia, Thessaloniki, Greece
The Rimini Centre for Economic Analysis (RCEA), Italy

AN OUT-OF-SAMPLE TEST FOR NONLINEARITY IN FINANCIAL TIME SERIES: AN EMPIRICAL APPLICATION

AN OUT-OF-SAMPLE TEST FOR NONLINEARITY IN FINANCIAL TIME SERIES: AN EMPIRICAL APPLICATION

Theodore Panagiotidis
Department of Economics
University of Macedonia
Thessaloniki 54006
Greece
Email: tpanag@uom.gr

16th May 2010

Abstract

This paper employs a local information, nearest neighbour forecasting methodology to test for evidence of nonlinearity in financial time series. Evidence from well-known data generating process are provided and compared with returns from the Athens stock exchange given the in-sample evidence of nonlinear dynamics that has appeared in the literature. Nearest neighbour forecasts fail to produce more accurate forecasts from a simple AR model. This does not substantiate the presence of in-sample nonlinearity in the series.

JEL C22, C53, G10
Keywords: nearest neighbour, nonlinearity

Acknowledgments: The author would like to thank but not complicate in any way David Chappell, Terence C. Mills and an anonymous referee.

1. INTRODUCTION

The growing interest in the application of nonlinear dynamics to a variety of physical and social science issues has been a significant theme in research over the last few years (see Granger 2008 for instance). Two additions to standard testing and estimation are crucial in the study of dynamic nonlinear models for financial time series. On the one hand, in-sample testing and, on the other, out-of-sample forecasting are fundamental in evaluating the reliability of nonlinear modelling results. The latter is especially true for nonlinear models where the threat of overfitiing is present. Since forecasting is an interdisciplinary subject, researchers regularly look to developments in other research areas. Progress in the physical sciences has been made in the area of using forecasts as a means of identifying non-linear deterministic components in a time series (Casdagli 1989, 1992). This involves evaluating time series predictability by using concepts, such as local nearest neighbours methods, that have demonstrated success in modelling non-linear deterministic data. In this paper we employ the forecasting test, which is an attempt to exploit residual dependence to improve forecasts of the level of the process. Success in forecasting demonstrates that residual dependence can be exploited to improve level specification.

The nonparametric, local information, forecasting technique, under consideration can be used to test for evidence of non-linear deterministic components in the underlying data generating process. One important reason to focus on this technique is that it has been demonstrated to do very well at forecasting non-linear deterministic systems. Casdagli (1992) reports that nearest neighbour techniques are able to forecast moderately high-dimensional deterministic non-linear functions extremely well out-of-sample. Thus, the results of nearest neighbour forecasting can be interpreted as a diagnostic for significant nonlinearity in financial data. A second reason for focusing on nearest-neighbour methods is their relation to the correlation-integral based tests for dependence. For example, the BDS test statistic (Brock et al 1996) is

$$W_{m,T}(\varepsilon) = \frac{\sqrt{T}\left\{C_{m,T} - C_{1,T}(\varepsilon)^{m}\right\}}{\sqrt{\sigma_{m,T}^{2}(\varepsilon)}} \qquad (1)$$

where $C_m(\varepsilon)$ denotes the fraction of *m*-histories in the series, which are within a distance of ε each other. The nearest-neighbour algorithm is essentially a method of systematically shifting through a data set looking for common/close *histories*. If the correlation-integral based test is, in fact, identifying important structural features of the data, then the nearest-neighbour algorithm ought to forecast the data very well.

Nearest-neighbour techniques have received attention in the literature lately. Applications using financial time series include LeBaron (1992), Mizrach (1992), Agnon et al (1999), Fernandez-Rodriguez and Sosvilla-Rivero (1998) while Jaditz and Sayers (1998), Jaditz et al (1998)and Golan and Perloff (2004) employ macroeconomic variables.

The rest of the paper is organised as follows. Section 2 provides the methodological framework. Two applications from a known generating process are presented in Section 3. Section 4 discusses the results. Last, section 5 concludes.

2. METHODOLOGY

In this application we follow the forecasting algorithm proposed by Casdagli (1992) and Jaditz and Riddick (2000). We start with data of the form $\{y_t, \mathbf{x}_t\}_{t=1,T}$, where y_t is a vector and $\mathbf{x}_t$ is a vector of conditioning information, where the elements of $\mathbf{x}_t$ are lags of the variable y_t. The time series are divided into two separate parts: a *fitting* set **F** and a *prediction* set **P**. We follow the usual practice of withholding the later observations that form the prediction set:

$$\mathbf{P} = \{(y_t, \mathbf{x}_t): N_f < t \leq T)\} \qquad t = N_{f+1}, \ldots, T$$

for some $N_f < T$.

In a univariate framework, we choose an embedding dimension m and construct a set of ordered pairs $\{(x_t, x_{t-1}^m)\}_{t=m+1,T}$, where x_t is the last available vector. The distance between x_{t-1}^m and x_{s-1}^m for all $s \in F$ is computed, for each x_t in the prediction set. The distances are ordered, we select the k nearest neighbours and fit a model of the following form

$$x_s = \alpha_{0,m,k} + x_{s-1}^m \alpha_{k,m} + \varepsilon_{s,k,m}$$

where the parameters $\alpha_{0,m,k}$ and $\alpha_{k,m}$ are estimated by ordinary least squares. The estimated parameters $\alpha_{0,m,k}$ and $\alpha_{k,m}$ are used to calculate the prediction

$$\hat{x}_t = \hat{\alpha}_{0,m,k} + x_{t-1}^m \hat{\alpha}_{k,m}$$

The prediction is then used to calculate the prediction error $x_t - \hat{x}_t = e_{t,m,k}$. These steps and the calculations are repeated for all the x's in the prediction set.

To calculate the distances, at least three alternatives have been suggested. Casdagli (1992) suggests using the sup norm to calculate distances,

$$\|x\| = \max_i |x_i|$$

Others (Cleveland and Devlin, 1988) advocate the Euclidean norm

$$\|x\| = \left(\sum_i x_i^2 \right)^{0.5}$$

or one could maximise the correlation function

$$\rho(x_i)$$

i.e. look for the highest serial correlation (Fernandez-Rodriguez and Sosvilla-Rivero 1998). In this case we have used the sup norm because it is less computationally expensive.

3. KNOWN DATA GENERATING PROCESSES

In this part, we follow Jaditz and Riddick (2000), who demonstrated the forecasting test in two known generating processes. Numerous numerical experiments by Casdagli (1992) showed that, for large data sets from non-linear deterministic data gerenerators, plots of the normalised RMSFE (Root Mean Square Forecast Error) achieve a very distinctive shape. The typical plot achieves a global minimum for a relatively small number of nearest-neighbours and is more or less continuously upward sloping as more nearest-neighbours are added, out to the limit at which we are essentially replicating the global linear predictor. As Casdagli (1992) concludes "*the use of graphical techniques should not be underestimated*". This distinctive shape will be demonstrated in the following example where this method is applied to data generated by the Henon system.

The Henon (1976) system is an example of a non-linear deterministic dynamical system. This system evolves perfectly deterministically from a given initial condition, in a pattern that appears highly erratic to the eye. It is given by the following pair of equations:

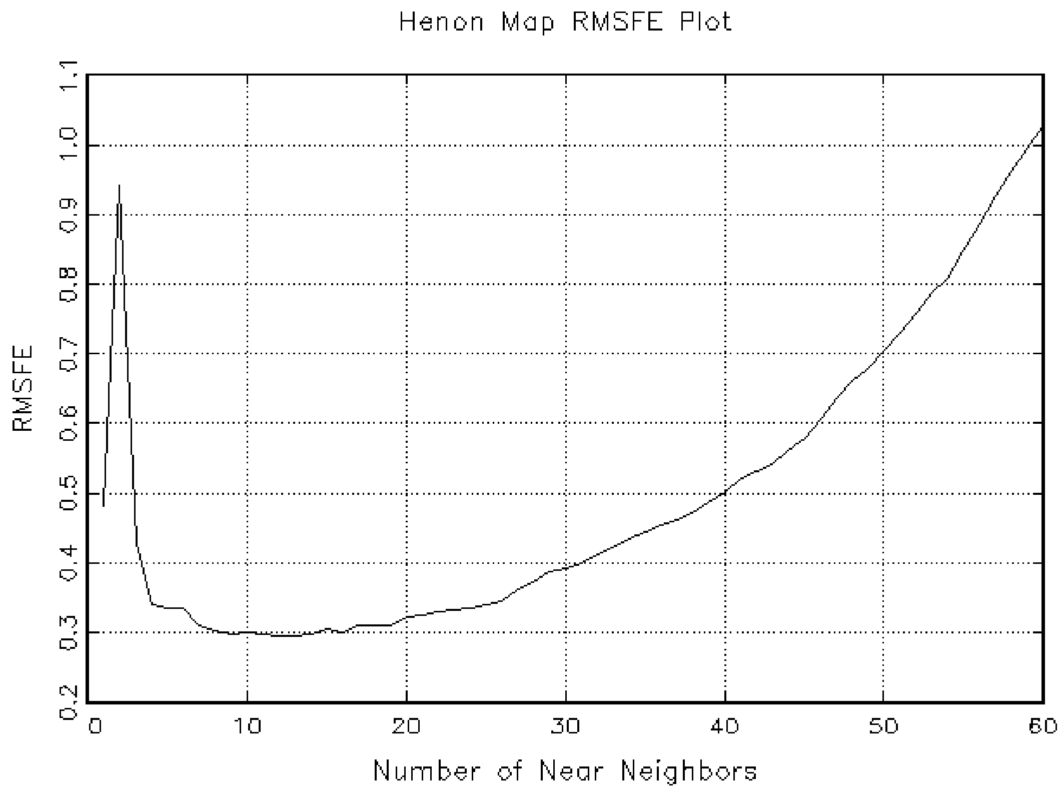
$$\begin{aligned} x_t &= y_{t-1} + 1 - 1.4x_{t-1}^2 \\ y_t &= 0.3x_{t-1} \end{aligned} \qquad (2)$$

The time paths of this system evolve on a fractal attractor that is known to have a dimension of 1.3. At embedding dimension equal to 3, the Henon map is perfectly forecastable.

We start with generating two data sets consisting of 100 observations. The first data set is generated by the Henon recursion using equation (2) and the second data set is comprised of output from the GAUSS standard normal pseudorandom number generator. We divide the two data sets into a fitting set with 60 observations and a fixed-window prediction set with 40 observations. The sup norm is used to calculate distances as suggested by Casdagli (1989). For each of the sample data sets, we estimate a naïve AR(1) model.

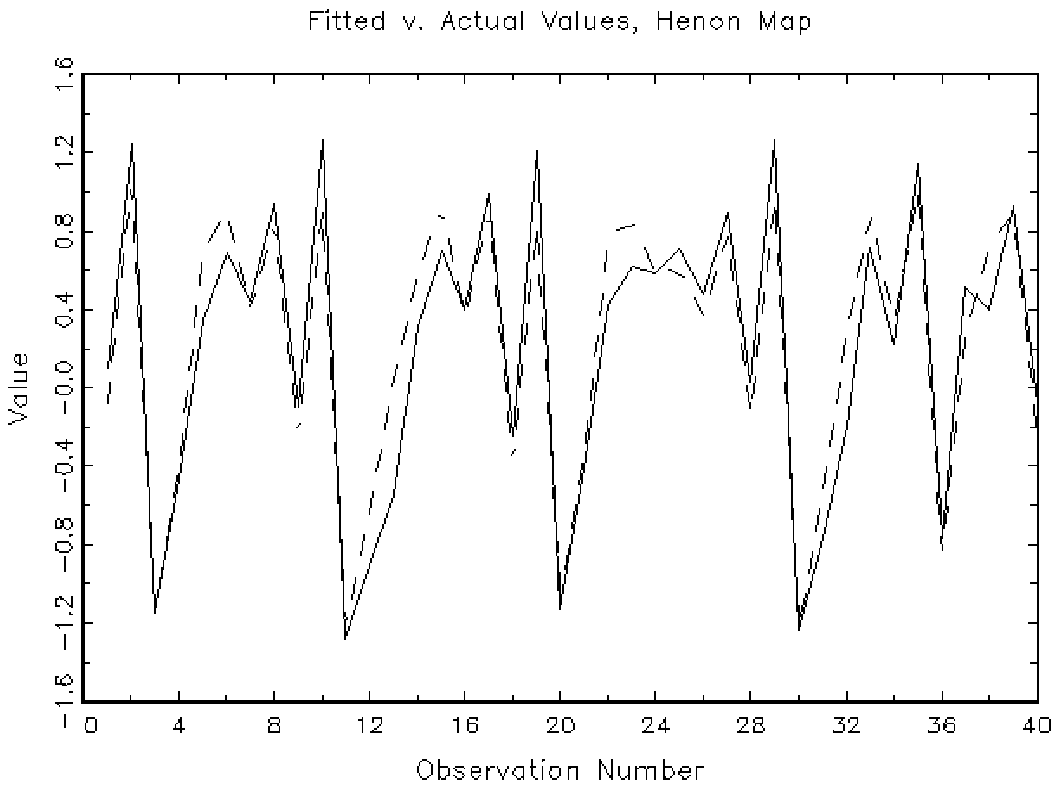
Figure 1 presents the plot the RMSFE as a function of the number of nearest neighbours included on the right-hand side of the regression equation. This picture exhibits two features that are highly characteristic of significant non-linearity in the underlying data generating process. First, the minimum point of the RMSFE plot occurs for a very small number of nearest neighbours. The best fitting regression uses only 13 nearest neighbours and has an RMSFE of 0.2943. Second, the forecasting performance deteriorates rapidly as more and more nearest neighbours are added to the regression. With this particular data set generated from a Henon system, the worst forecast are from the AR(1) model estimated using all 60 observations in the fitting set, which has an RMSFE of 1.026.

Figure 1: Henon Map, Number of Nearest Neighbours vs. RMSFE



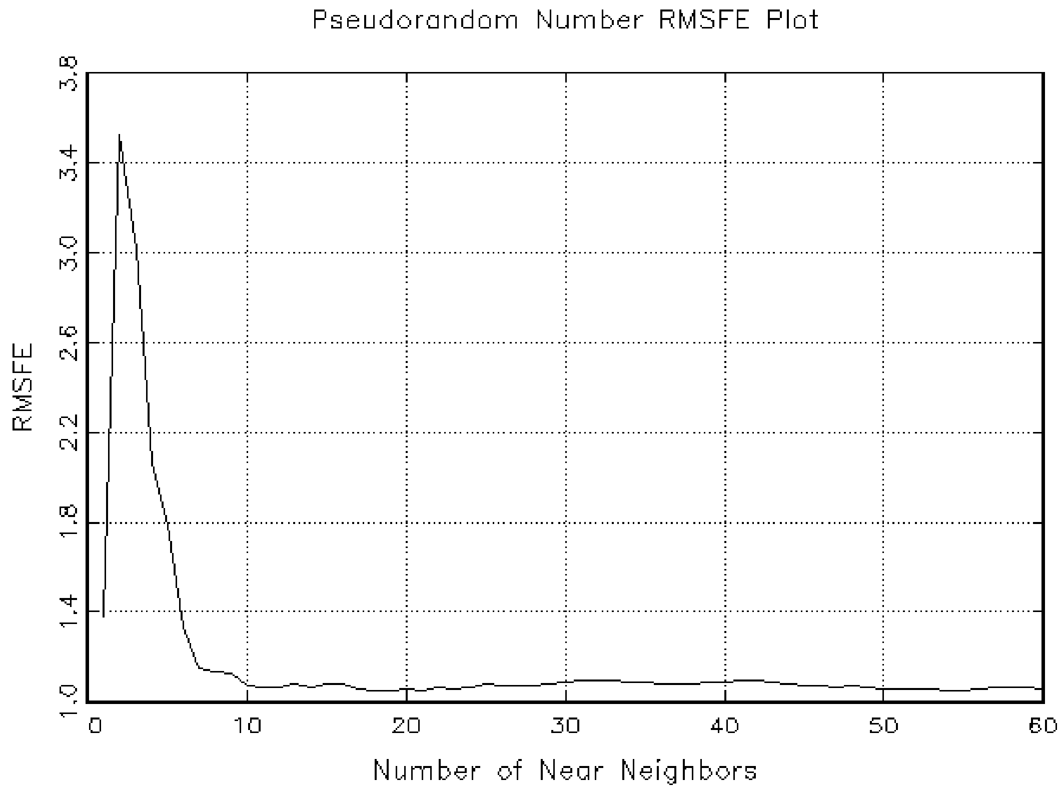
Given that the nearest-neighbour regression with 13 nearest neighbours (*nn*) produced the lowest RMSFE, we are going to use that to generate forecasts employing the observations that belong to the prediction set. Figure 2 is a plot of the actual versus the predicted values for the nearest-neighbour regression with 12 *nn*. This confirms that the best nearest-neighbour regression forecasts much of the variation of the prediction set.

Figure 2: Henon Map: actual vs. fitted values (dotted line is the predicted one)



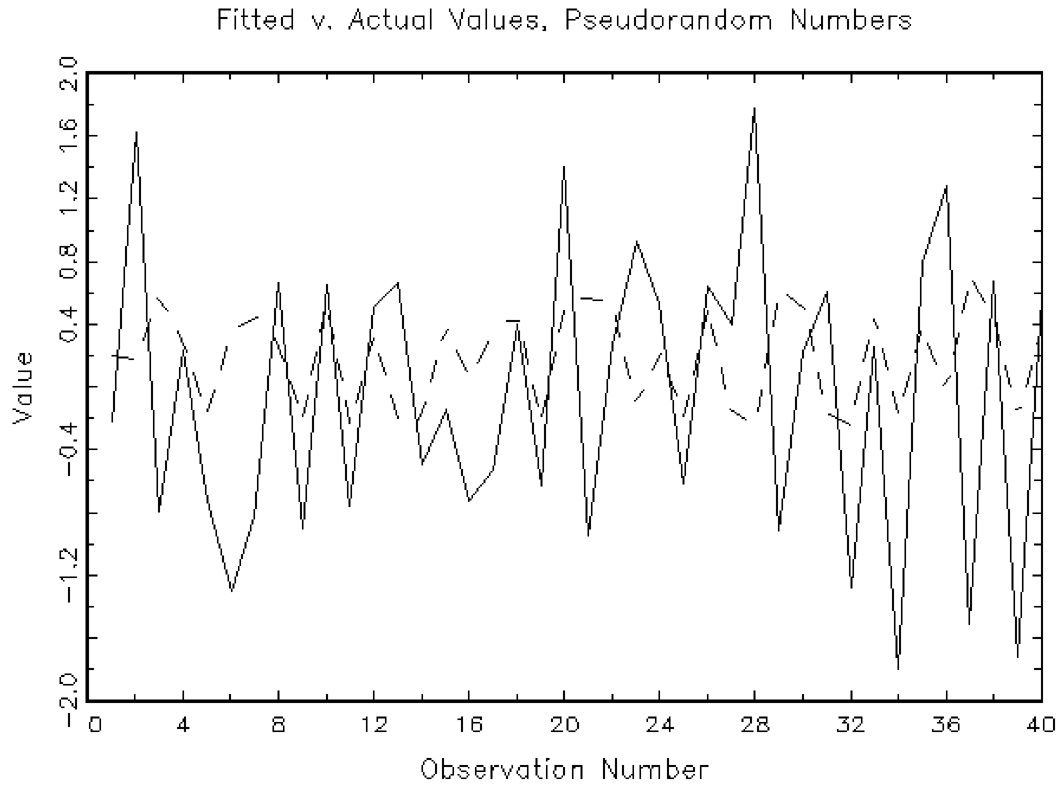
The same methodology is followed in the case of the second data set produced from the random data generator. Figure 3 presents the RMSFE plot against the number of *nn* for the Gaussian pseudorandom numbers data set. Although there is a sharp peak in the RMSFE plot for a very small number of nearest-neighbours, one could observe that the plot is essentially flat. Indeed, if the true model is linear, one could argue that the RMSFE should be downward sloping, with forecast accuracy improving at rate $\sqrt{n}$, where *n* is the number of nearest-neighbours used to estimate the regressions. The minimum of the RMSFE plot occurs at 18 *nn*, with RMSFE of 1.0449

Figure 3: Pseudorandom numbers: number of nearest neighbours vs. RMSFE



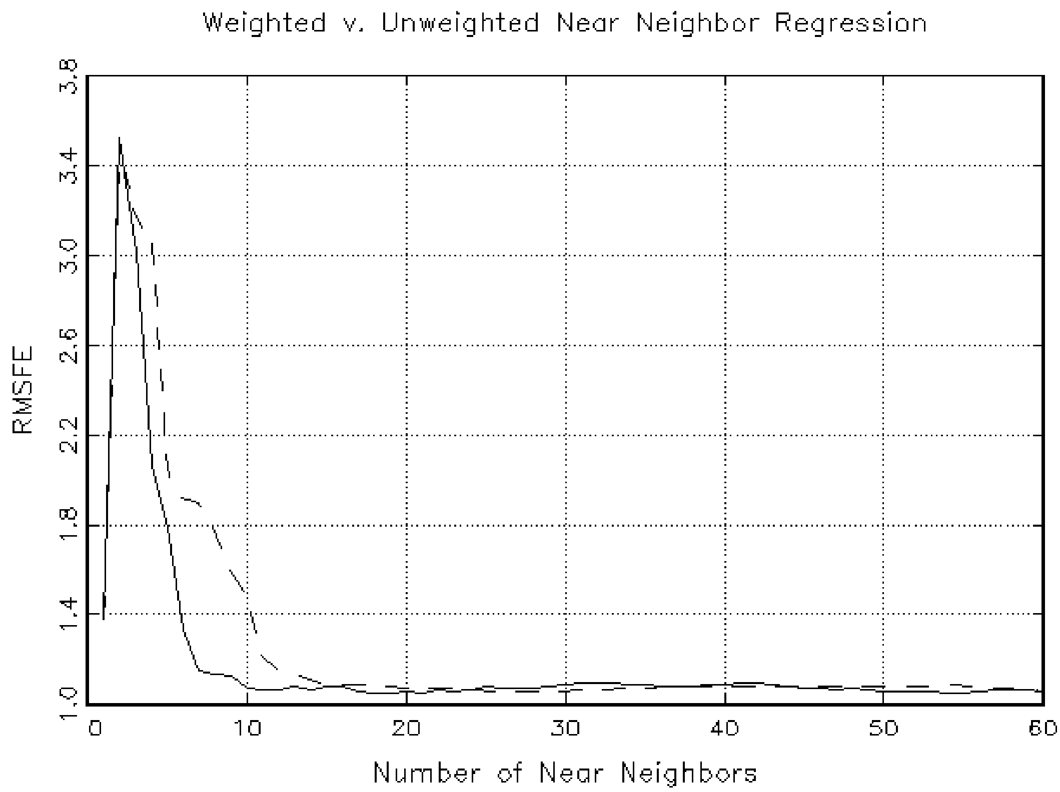
As noted earlier, the lowest RMSFE *nn* regression (*n*=18) was used to forecast the prediction set. This is presented in Figure 4. The plot of actual versus predicted values confirms the lack of fit and underlines the fact that the *nn* are unable to generate superior forecasts.

Figure 4: Pseudorandom numbers, actual versus fitted values (the dotted line is the predicted one)



Additionally, we compared the forecasting ability of (unweighted) simple OLS *nn* AR(1) regression with the local weighting schemes suggested by Cleveland & Devlin (1988) that place greater weights on nearby observations in estimating the local linear regression. The tricube function is used to calculate the weights for the weighted least-squares parameter vector as proposed by Cleveland & Devlin (1988). The advantage of the latter (weight the "closest" observations) does not come without cost (e.g. speed). Figure 5 plots the RMSFE curves for both the weighted and the simple OLS *nn* AR(1) regression, estimated on the Gaussian pseudorandom number data. For 40 of the 60 regressions, the unweighted *nn* routine has a lower RMSFE than the weighted *nn* regression. Summing up, we find that it is more typically the case that the unweighted algorithm outperforms the weighted regression.

Figure 5: Random Numbers: weighted vs. unweighted regression (dotted line)



All the above examples, illustrate two points. First, it is possible that the local-information forecasting algorithm can yield large improvements in forecast accuracy. Second, these examples illustrate that the minimum point of the RMSFE plot occurs for a small number of *nn*. In all cases, prediction performance degrades smoothly as more and more *nn* are added to the local regression.

To conclude, these results illustrate how the nearest-neighbours methodology may be useful in identifying whether evidence of deterministic non-linear dynamics is present in the level equation of an unknown data generator. The RMSFE plots could be a very useful informal diagnostic. The important feature of these plots is that, when nonlinearities are present, the plots appear to have a distinctive upward slope. If the data generator is deterministic, the slope may be quite striking. For stochastic data generators, the slope is typically more shallow. Of course, "real data" will fall between these two extremes. If the data generator is "linear", then we expect to see fairly flat RMSFE plots similar to the one in Figure 3. On the other hand, if the data generator is "non-linear", we would expect to see upward-sloping plots with a well-defined minimum. With enough observations, we may be able to reject the null of equal forecast accuracy between a global information linear forecast and the best "local" information non-linear forecast. Casdagli (1992), Jaditz & Sayers (1998) and Jaditz & Riddick (2000) provide numerous examples calculated on deterministic data generators which further illustrate this point.

4. EVIDENCE FROM FINANCIAL TIME SERIES

Significant in-sample non-linearities were uncovered in the General Index of the Athens Stock Exchange (ASE) in Panagiotidis (2010), in which a battery of *iid* tests were employed including the Brock et al (1996) (BDS), McLeod and Li (1983), Engle (1982), Tsay (1986) tests and the Bicovariance Test (Hinich and Patterson 1995). Chappell and Panagiotidis (2005) further investigate the nonlinearities using the correlation dimension. Given the significant nonlinearities found in these studies, we would expect our robust methodology to be able to offer a significant forecast improvement. This paper is going one step further as it is one of the first papers to employ the forecasting test using financial time series that are known to be nonlinear.

The data are daily returns of the General Index of the Athens Stock Exchange, calculated from daily closing prices, and the sample period is from 1st June 2000 to 31st December 2002. Unit roots confirm that the returns are stationary (Panagiotidis 2010).

Figure 6 presents the results. The minimum RMSFE occurs at 269 nearest neighbors, where RMSFE = 0.9690. Using the *nn* that minimised the RMSFE we produced forecasts for the last 50 observations with fixed window fitting set (the same was used in the known data generators in the previous section). Three characteristics emerge from our analysis. First, the RMSFE plot is flat and not upward sloping, implying that non-linearities may not be present in the mean equation. Secondly, the minimum point of the RMSFE plot does not occur for a very small number of *nn*. Thirdly, not surprisingly, the forecasting exercise is not successful. The *nn* forecasts fail to capture the variability of the series (see Figure 7).

Lastly, in order to confirm the results of the previous section we present the weighted vs. the unweighted regression. The weighted regression does not improve the forecasting ability on the one hand and on the other increases the computational time considerably (see Table 1 and Figure 8).

Figure 6: ASE General Index: number of *nn* vs. RMSFE

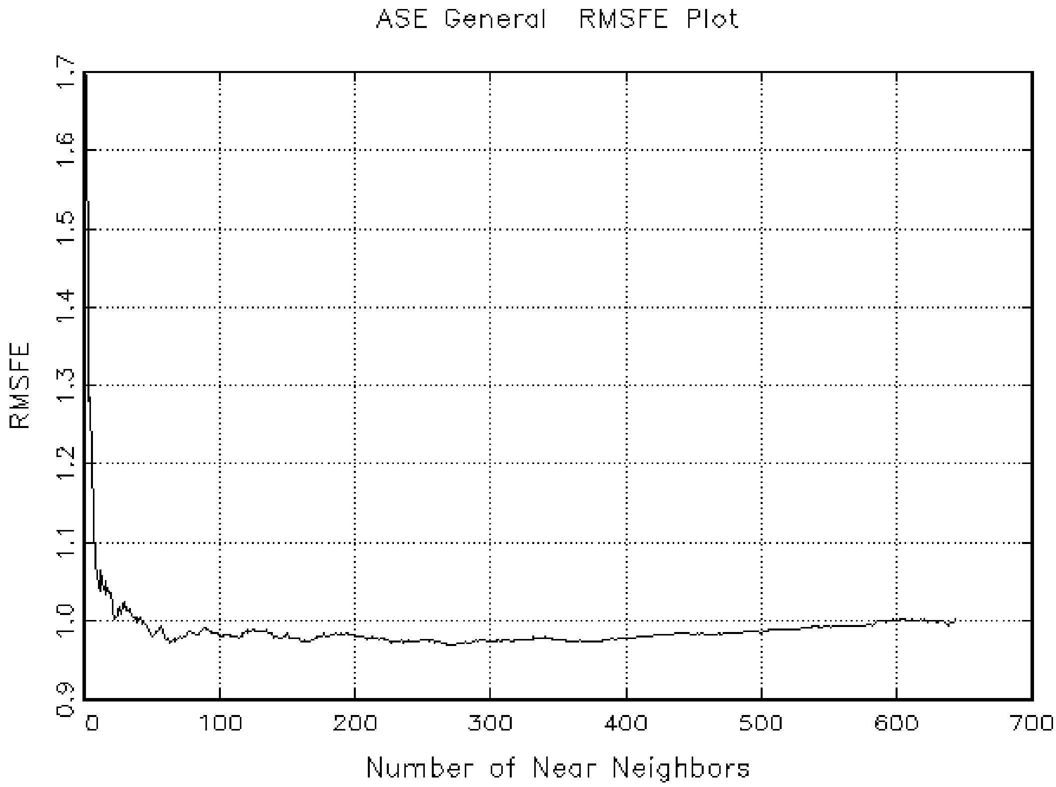


Figure 7: ASE General Index actual vs. fitted values

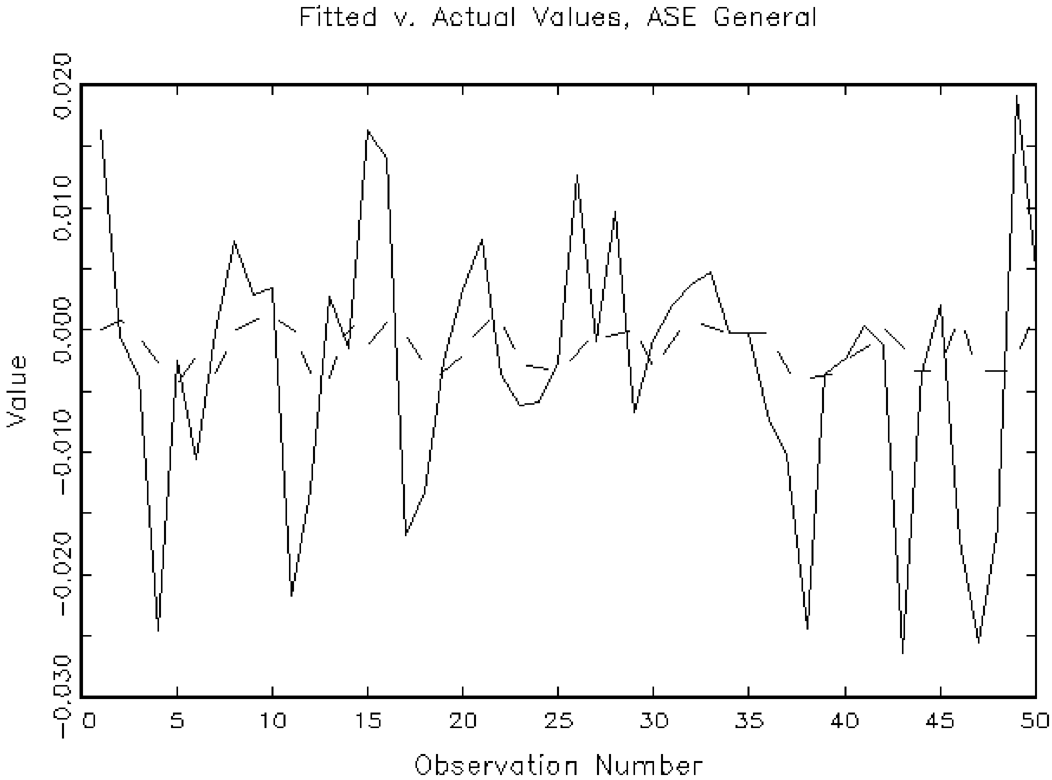


Figure 8: Weighted vs. Unweighted Nearest Neighbour Regression (the weighted regression is represented by the dotted line)

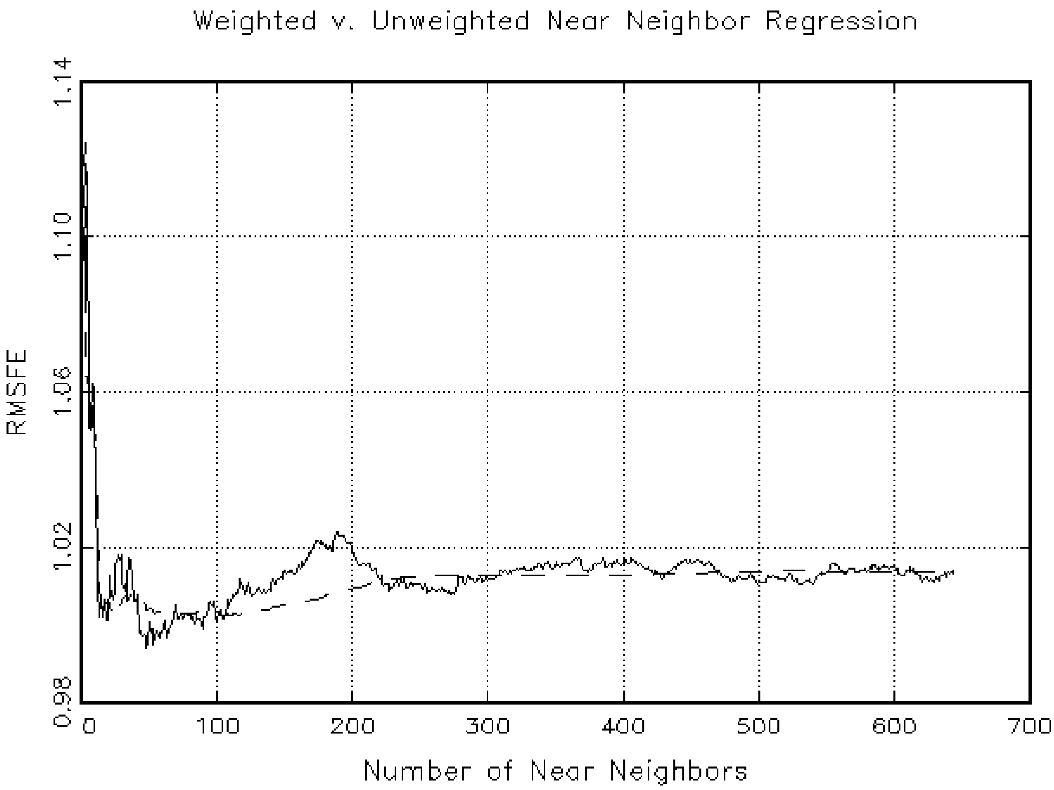


Table 1

First 10 forecast RMSFE's:		
NB: For k=1, we use the simple local mean forecast.		
k	RMSFE Unweighted	RMSFE Weighted
----	----------------	----------------
1	1.097	1.097
2	1.091	1.124
3	1.084	1.063
4	1.117	1.064
5	1.051	1.057
6	1.050	1.057
7	1.053	1.056
8	1.062	1.057
9	1.061	1.045
10	1.031	1.049

It would be useful to ask whether the results of our exercise are sensitive to our assumptions. To answer this questions we have repeated the exercise, which takes approximately 45 minutes in a Pentium 4 PC. This time we replaced the fixed window fitting set with a sliding window, where one predicts using only the most recent observations. Additionally, we have expanded the prediction set from 50 in the previous example to 150. The results are presented in Figures 9, 10 and 11 and Table 2.

The RMSFE plot is presented in Figure 9 and the minimum RMSFE occurs at 280 nearest neighbors, where RMSFE = 0.9876 (compared with 269 *nn* and 0.969 in the previous application). Again, none of the characteristic features appear and the *nn* that minimises the RMSFE is unable to generate meaningful forecasts as it fails to capture the volatility of the series (Figure 10). Lastly, our conclusions with regard to the weighted regression is confirmed in this case as it is difficult to differentiate between the latter and the unweighted regression (see Figure 11 and Table 2)

Figure 9

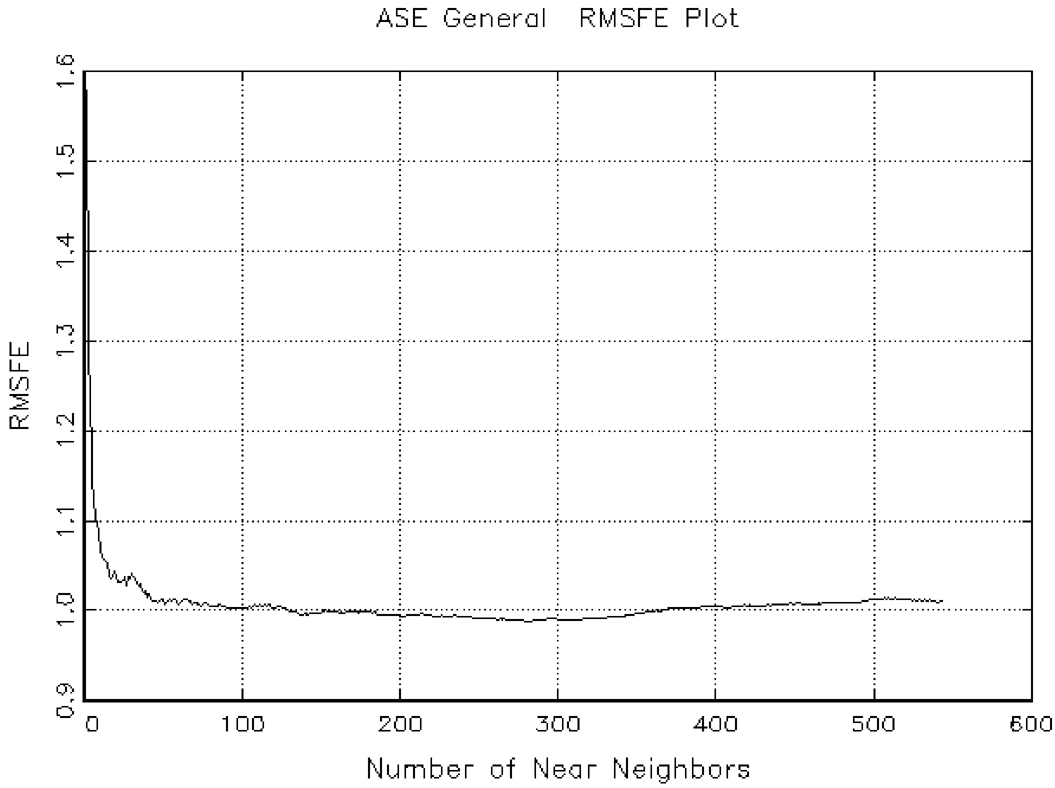


Figure 10

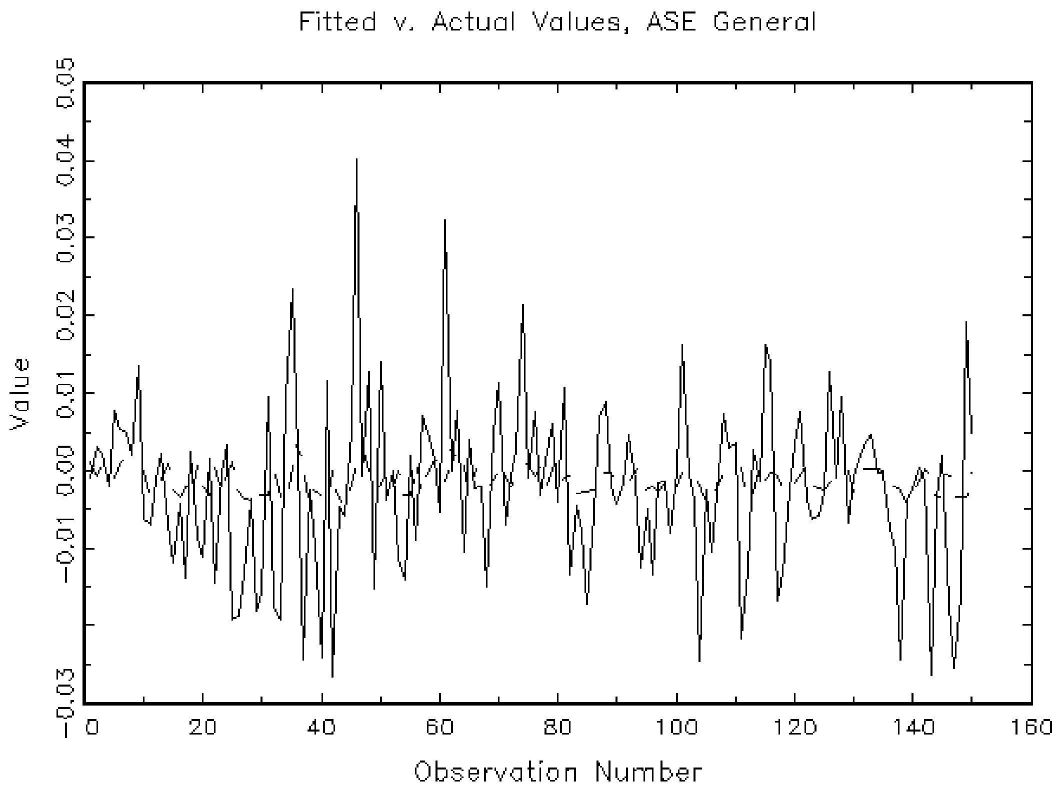


Figure 11

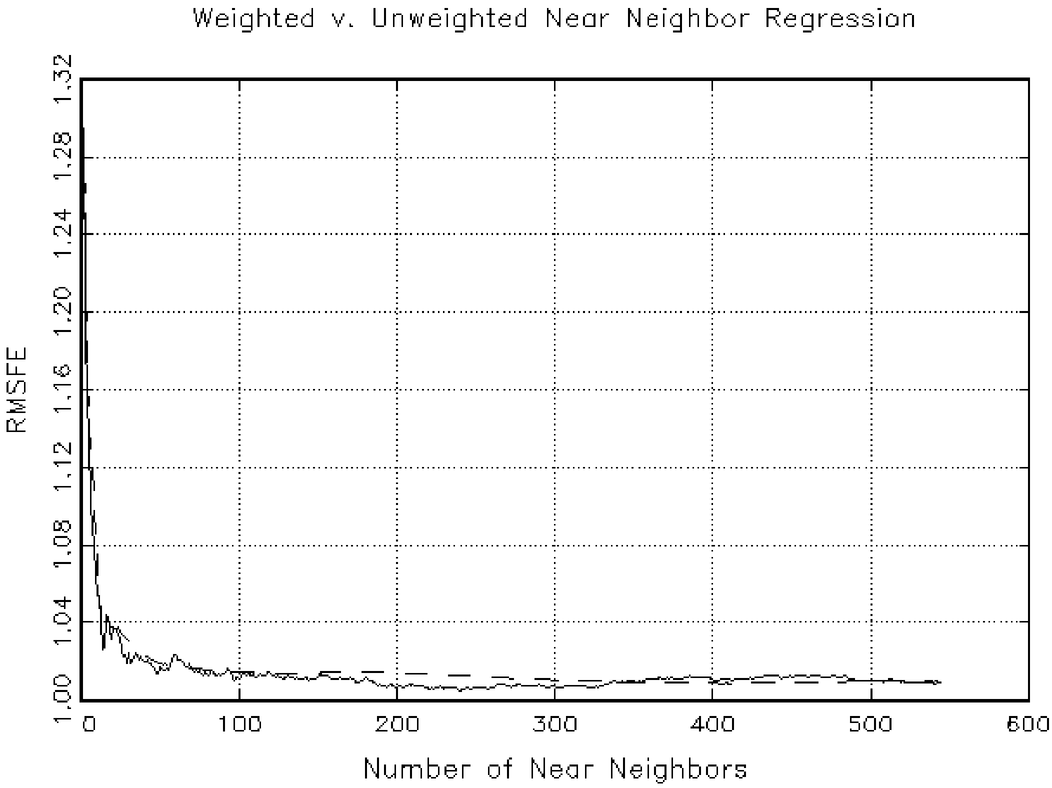


Table 2

First 10 forecast RMSFE's:
NB: For k=1, we use the simple local mean forecast.

k	RMSFE Unweighted	RMSFE Weighted
1	1.294	1.294
2	1.201	1.241
3	1.146	1.191
4	1.145	1.157
5	1.137	1.138
6	1.101	1.128
7	1.091	1.112
8	1.079	1.100
9	1.066	1.082
10	1.055	1.061

5. Conclusions

This paper employed a local information, nearest-neighbour forecasting methodology to test for evidence of nonlinearity in financial time series. Returns from the Athens Stock exchange were investigated given the in-sample evidence of nonlinear dynamics that has appeared in the literature recently. Evidence from well-known data generating processes are provided and compared with the ASE returns. We fail to find nearest neighbour forecasts that are significantly more accurate than forecasts from simple AR models. Our results fail to substantiate the presence of in-sample nonlinearity in the series.

REFERENCES

Agnon, Y., Golan, A., Shearer, M., (1999), Nonparametric, nonlinear, short term forecasting: theory and evidence for nonlinearities in the commodity markets, *Economic Letters*, **65**, 293-299.

Brock, W.A., Dechert, W., Scheinkman J. and LeBaron, B. (1996), A Test for Independence based on the Correlation Dimension, *Econometrics Reviews,* **15**, 197-235.

Casdagli, M. (1989), Nonlinear prediction of chaotic time series, *Physica D*, **35**, 335-356.

Casdagli, M. (1992), Chaos and Deterministic versus Stochastic Non-Linear Modelling, *Journal of the Royal Statistical Society*, Series B (Methodological, 54 (2), 303-328.

Chappell, D. and Panagiotidis, T. (2005), Using the correlation dimension to detect non-linear dynamics: Evidence from the Athens Stock exchange, *Finance Letters*, **3** (4), 29-32.

Cleveland, W.S. and Delvin, S.J. (1988), Locally weighted regression: An approach to regression analysis by local fitting, Journal of the American Statistical Association, **83**, 596-610.

Engle, R.F. (1982), Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation, *Econometrica*, **50**, 987-1007.

Fernandez-Rodriguez, F., Sosvilla-Rivero, S., (1998), Testing nonlinear forecastability in time series: Theory and Evidence from the EMS, *Economics Letters*, **59**, 49-63.

Golan, A. and Perloff J.M. (2004), Superior Forecasts of the U.S. Unemployment rate using a nonparametric method, *Review of Economics and Statistics*, **86**, 1, 433-438.

Granger, C.W.J. (2008), Non-Linear Models: Where Do We Go Next - Time Varying Parameter Models? *Studies in Nonlinear Dynamics & Econometrics*, **12(3).**

Henon, M. (1976), A two dimensional mapping with a strange attractor, *Communications in Mathematical Physics*, **50**, 69-77.

Hinich, M.J. and Patterson, D.M. (1995), Detecting Epochs of Transient Dependence in White Noise, unpublished manuscript, University of Texas at Austin.

Jaditz, T., and Riddick, L., (2000), Time-Series Near-Neighbor regression, *Studies in Nonlinear Dynamics and Econometrics*, **4** (1), 35-44.

Jaditz, T., Riddick, L.A., and Sayers, L., (1998), Multivariate Nonlinear forecasting, *Macroeconomic Dynamics*, **2**, 369-82.

Jaditz, T., Sayers, C.L., (1998), Out-of-Sample Forecast Performance as a test for nonlinearity in time series, *Journal of Business and Economic Statistics*, **16** (1), 110-117.

LeBaron, Blake, (1992), Forecast Improvements Using a Volatility Index, *Journal of Applied Econometrics*, 7, 137-149.

McLeod, A.I. and Li., W.K. (1983), Diagnostic Checking ARMA Time Series Models Using Squared-Residual Autocorrelations, *Journal of Time Series Analysis,* **4**, 269-273.

Mizrach, B., (1992), Multivariate nearest-Neighbour Forecasts of EMS Exchange Rates, *Journal of Applied Econometrics*, 7, 151-163.

Panagiotidis T. (2010), Market Efficiency and the Euro: The case of the Athens Stock Exchange, forthcoming *Empirica*.

Tsay, R.S. (1986), Nonlinearity tests for Time Series, *Biometrica*, **73**, 461-466.